Towards the Algorithmic Detection of Artistic Style

Jeremiah W. Johnson
Department of Applied Sciences & Engineering
University of New Hampshire
Manchester, NH 03101

***Abstract*—The artistic style of a painting can be sensed by the average observer, but algorithmically detecting a painting's style is a difficult problem. We propose a novel method for detecting the artistic style of a painting that is motivated by the neural-style algorithm of Gatys et al. and is competitive with other recent algorithmic approaches to artistic style detection.**

***Keywords*—Artificial intelligence; neural networks; style transfer; representation learning; deep learning; computer vision; machine learning**

I. Introduction

Any observer can sense the artistic style of painting, even if it takes formal training to articulate it. However, artistic style in general is not a well-defined concept; rather, it is loosely defined as "...a distinctive manner which permits the grouping of works into related categories" [1]. Although a vaguely defined concept, artistic style is often still the primary means used by art historians to classifying paintings, despite efforts in recent years by some experts to move away from a style-based classification toward a geographic and period-based classification instead [2].

Given the imprecise definition and the often limited number of examples of paintings in a particular style, algorithmically detecting the artistic style of a painting can be a challenging problem. The challenge is often compounded by the digitization process, which itself has consequences that may affect the ability of a machine to correctly detect artistic style; for instance, textures may be affected by the resolution of the digitization, and shadows created by external objects may occlude portions of the image. Despite these challenges, intelligent systems for detecting artistic style could be useful for a variety of applications, such as recommendation systems.

In recent years convolutional neural networks have been used to achieve remarkable results on a wide range of challenging tasks in computer vision, including object detection, semantic segmentation, instance segmentation, and image style transfer [3]–[6]. In this paper we build off of recent work of Gatys et al. using convolutional neural networks for image style transfer to develop a novel algorithm for artistic style detection.

The contribution of this paper is as follows:

- We propose a novel method for algorithmically detecting the artistic style of a painting. This method is motivated by the so-called neural style algorithm of Gatys et al. [5], and utilizes the Gram matrices of filter activations at early layers in a convolutional neural network to construct a learned representation that captures relevant information about stylistic aspects of the painting, such as style and color, while discarding information about image content.
- We demonstrate that our proposed method achieves competitive results when compared with other neural network based algorithms for artistic style detection, even when using a larger than typical number of artistic style categories, and we consider avenues for further improvement.

II. Related Work

A. Algorithmic Style Detection

The algorithmic detection of artistic style in paintings has only been considered sporadically in the past. Examples of early efforts at style classification are [7] and [8]. In these early examples, the datasets used are quite small and only a handful of very distinct artistic style categories considered.

More recently, Salah and Elgammal constructed several complex and effective models for artistic style detection using a variety of techniques including metric learning, feature fusion, and metric fusion [9]. These models rely on the incorporation of carefully hand-engineered and selected features. Although not primarily based on convolutional neural networks, these models do incorporate the learned representation from the last layer of a convolutional neural network that has been pretrained for image classification. The work of Salah and Elgammal uses a dataset similar in scale to the one used for this work.

One of the first examples of the use of convolutional neural networks for image style detection is [10]. In this work, convolutional neural networks that have been pretrained for image classification on the ImageNet dataset are finetuned to algorithmically detect image style, providing an early example of transfer learning with convolutional neural networks. This work primarily investigates the algorithmic detection of the style of photographic images, though it does include a brief investigation into algorithmic detection of artistic style in paintings. In both [9] and [10], the artistic style detection problem has been simplified somewhat by holding the number of artistic style categories to 25 and 27 respectively.

B. Neural Networks and Learning Representations

The intuitive explanation often given for the many recent successes of deep neural networks on challenging tasks in computer vision and natural language processing is that deep neural networks learn 'good' representations of the data that they are trained on [11], [12]. There is a lack of grounded theory on what exactly constitutes a 'good' representation of a given dataset, to the point that it is not clear that two identical

Fig. 1. Two examples of image style transfer generated using the neural style algorithm of Gatys et al. On the left is the content image, in the center, the style image, and on the right, the image generated via the 'neural-style' algorithm.

networks with different initializations will learn the same or even similar representations [12]. Despite the limited theory underpinning their use, the learned representations of neural networks have been used effectively for a wide range of tasks in recent years, including, in the computer vision domain, image classification, object detection, semantic segmentation, instance segmentation, and style transfer, among others. For instance, to perform instance segmentation of natural images, the state of the art Mask-RCNN model first extracts the learned representations of the data at various layers in a so-called 'backbone' convolutional neural network that has typically been pretrained for image classification. The model then uses these representations as input into several smaller convolutional neural networks [4].

Although there may be a lack of grounded theory on what a 'good' representation of a given dataset is, there is a significant body of recent research into the inner workings of convolutional neural networks that provides some insight into what the various layers of a convolutional neural network learn [13], [14]. In general, a convolutional neural network trained for image classification on a dataset such as ImageNet will learn a hierarchical set of representations of the training data, where the learned representations at lower layers in the model capture low-level aspects of the images, such as the presence of vertical or horizontal lines or other patterns, colors, and textures, while higher layers in the model learn more complex representations that are more aligned with the content of the image; c.f. Fig. 2. This suggests that the lower-level features in a convolutional neural network learn information relevant to an algorithmic determination of artistic style.

However, the raw lower-level features in a convolutional neural network turn out to not necessarily be the ideal features to use when considering artistic style. In the paper "A Neural Algorithm of Artistic Style", Gatys et al. demonstrated that by focusing not simply on the low-level feature activations in a convolutional neural network, but rather on the correlations between the low-level feature activations in a convolutional neural network, one could obtain significant useful information about low-level global properties of the image, such as its color and texture. This in turn enables the transfer of these aspects of the input image onto another image via an algorithm informally referred to as the 'neural-style' algorithm [5]. Two examples of the output of this algorithm are presented in Fig. 1. Since the paper of Gatys et al., in recent years several authors have built upon their work [6], [15], [16]. In [15] and [16], the investigations are focused on ways to improve either the quality of the style transfer. In [6], the algorithm is modified to produce a dramatic increase in the efficiency neural style transfer. To the best of our knowledge the only other work to consider the use of the style representation of an image for algorithmic style detection is [17]. In this work, the authors take an approach similar to that taken here, but with a much smaller dataset, a much smaller set of artistic style categories, and without comparison to other recent deep neural network based approaches.

III. Background: The Neural Style Algorithm

As described above, the primary insight in the neural-style algorithm outlined by Gatys et al. is that to some degree the correlations between low-level feature activations in a convolutional neural network capture important information about the style of the image, while higher-level feature activations capture information about the content of the image. Thus, to construct an image $\mathbf{x}$ that merges both the style of an image $\mathbf{a}$ and the content of an image $\mathbf{p}$, an image is initialized as white noise and the following two loss functions are simultaneously

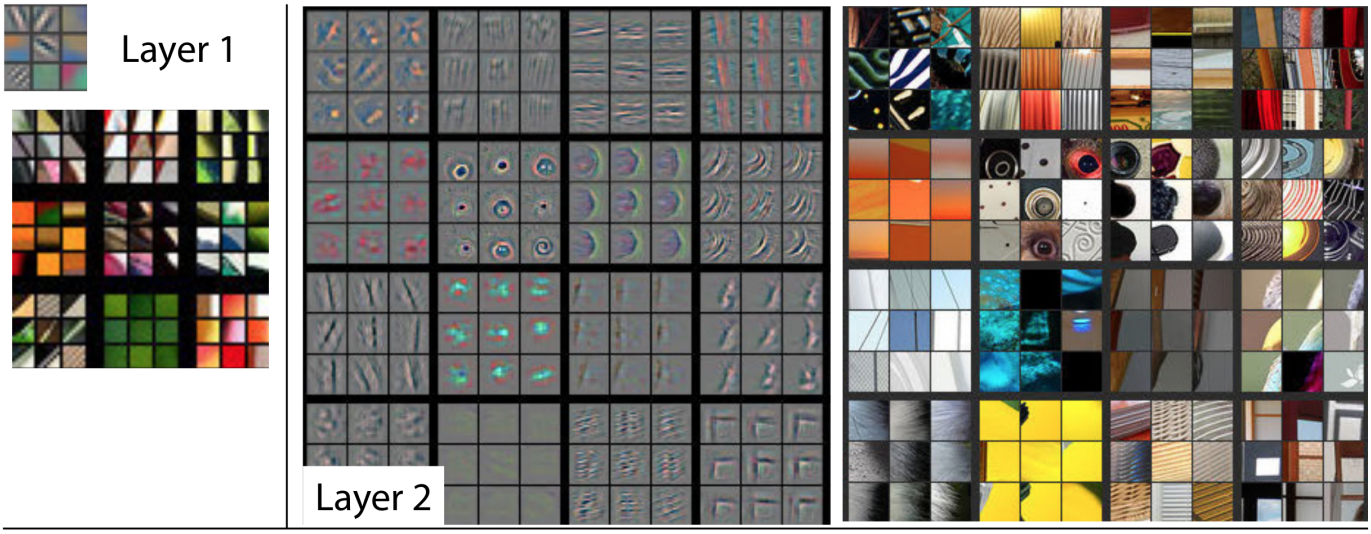

Fig. 2. A visualization of some of the learned features in the lower layers in AlexNet. Image from [13].

minimized:

$$\mathcal{L}_{content}(\mathbf{p}, \mathbf{x}) = \sum_{l \in L_{content}} \frac{1}{N_l M_l} \sum_{i,j} \left(F_{ij}^l - P_{ij}^l\right)^2, \quad (1)$$

and

$$\mathcal{L}_{style}(\mathbf{a}, \mathbf{x}) = \sum_{l \in L_{style}} \frac{1}{N_l^2 M_l^2} \sum_{i,j} (G_{ij}^l - A_{ij}^l)^2. \quad (2)$$

Here N_l represents the number of filters in layer l, M_l represents the spatial dimensionality of the feature map, $\mathbf{F}^l$ and $\mathbf{P}^l$ represent the feature maps extracted by the network at layer l from the images $\mathbf{x}$ and $\mathbf{p}$ respectively, and, letting $\mathbf{S}^l$ represent the feature maps extracted by the network at layer l from the image $\mathbf{a}$,

$$G_{ij}^l = \sum_{k=1}^{M_l} F_{ik}^l F_{jk}^l, \text{ and } A_{ij}^l = \sum_{k=1}^{M_l} S_{ik}^l S_{jk}^l. \quad (3)$$

The style loss function $\mathcal{L}_{style}$ above is the component of the model responsible for capturing relevant style information from image $\mathbf{a}$ and transferring it into image $\mathbf{x}$. As can be seen in 3, this loss is calculated over the Gram matrices of the filter activations at the specified layers.

IV. Data and Methods

A. Data

The data used for this investigation consists of 76449 digitized images of fine art paintings. The majority of the images were originally obtained from http://www.wikiart.org, which is currently the largest online repository of fine-art paintings. For convenience, a prepackaged imageset sourced and prepared by Kiri Nichols and hosted by the data-science competition website http://www.kaggle.com was used for the experiments documented in this paper. A stratified 10% of the dataset was held out for validation purposes. A more fine-grained set of style categories for classification than has been used in previous work on artistic style detection was chosen, as finer classification is likely necessary for practical application. For the experiments described in this work, 70 distinct style categories are used, the maximum amount possible with the current dataset while insuring the existence of at least 100 observations in each style category. This noticeably increases the complexity of the classification task, as many of the class boundaries are not well-defined, the classes are significantly unbalanced, and there are not nearly as many examples of each of the artistic styles as in previous work on large-scale algorithmic artistic style detection.

TABLE I. Baseline Results

Model	Accuracy (top 1%)
Convolutional Neural Network	27.47
Pretrained Residual Neural Network	36.99

B. Baseline Models

To establish a baseline for algorithmic artistic style detection, a single convolutional neural network was first trained from scratch. The network has a uniform structure consisting of convolutional layers with 3x3 kernels and leaky ReLU activations ($\alpha = 0.333$). Between every pair of convolutional layers is a fractional max pooling layer with a 3x3 kernel. Fractional max-pooling is used as given the relatively small size of the dataset, the more commonly used average or max-pooling operations would lead to rapid data loss and a significantly more shallow network [18]. The convolutional layer sizes are $3 \rightarrow 32 \rightarrow 96 \rightarrow 128 \rightarrow 160 \rightarrow 192 \rightarrow 224$, followed by a fully-connected layer and 70-way softmax. We applied 10% dropout to the fully connected layer. Aside from mean normalization and horizontal flips, the data were not augmented in any way. The model was trained over 55 epochs using stochastic gradient descent with a learning rate of 0.1

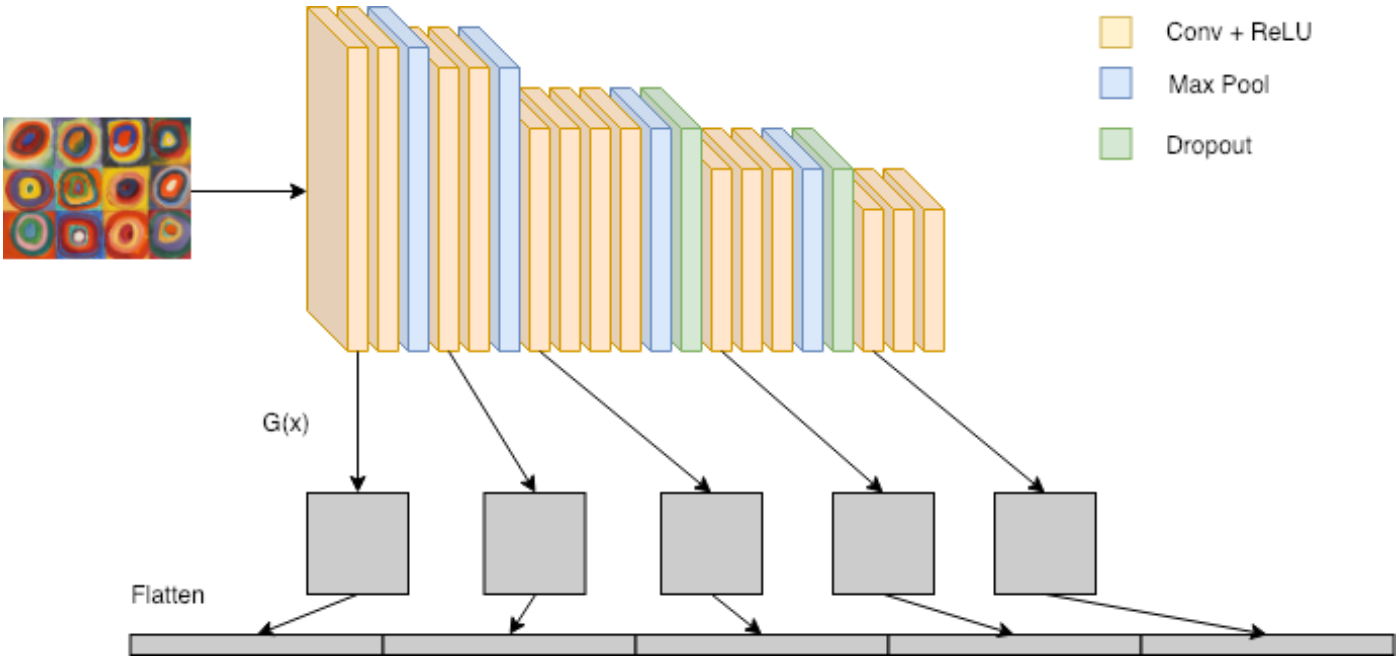

Fig. 3. Illustration of the construction of a neural style representation using a VGG16 style network. $G(x)$ denotes the Gram matrix calculation.

and achieved a top 1% accuracy of 27.468%.

We then finetuned a pretrained image classification model for algorithmic artistic style detection. The model used was a residual neural network with 50 layers pretrained on the ImageNet 2015 dataset [19]. There are two motivating factors for choosing to finetune this network. The first is that residual networks currently exhibit the best results on image classification tasks, and previous work on algorithmic detection of artistic style suggests that a network trained for the task of image classification and then finetuned for artistic style detection is likely to perform the task well [20], [10]. The second and more interesting reason from the standpoint of artistic style classification is that the architecture of a residual neural network makes the outputs of lower levels of the network available unadulterated to higher levels in the network. In this way, residual networks have been noted to function similar to a Long Short-Term Memory network without gates [21]. For style classification, this is particularly appealing as a means of allowing the higher levels in the net to consider both lower-level features and higher-level features when forming an artistic style classification, where the style may very much be determined by the lower-level features. The finetuned residual neural network model obtained top-1% accuracy of 36.985%. Results for the baseline models are summarized in Table I.

C. Neural Style Representation Models

To construct the neural style representation for use in algorithmic artistic style detection, we extracted the feature activations at layers ReLU1_1, ReLU2_1, ReLU3_1, ReLU4_1, and ReLU5_1 from the nineteen-layer convolutional neural network developed by the Visual Geometry Group at the University of Oxford, the so-called VGG-19 model, for the paintings described above [22]. We then calculate the Gram matrices of these activations. The model and layers used were chosen based on the quality of the style transfers obtained by [5] using this network and layers, while the weights for the pretrained VGG-19 model was obtained from the Caffe Model Zoo [23]. The Gram matrices were then reshaped to account for symmetry, producing a total of 304,416 distinct features per image. This process is illustrated in Fig. 3.

Algorithmic style detection via the style representation was approached in two ways. First, the full feature vector was normalized and then passed to a single-layer linear classifier. This classifier was trained online using the Adam optimization algorithm for 55 epochs, and achieved a top 1% accuracy of 13.23%. [24]. It should be noted that the online training approach taken here was likely not optimal, and was dictated by the high dimensionality of the data and corresponding hardware constraints, which in turn limited the batch size and the hyperparameter search space.

Next, to help mitigate the previously mentioned dimensionality constraints of the full neural style representation, we investigated the representations learned at each layer individually. After extracting the Gram matrices at each of the five layers mentioned above, we built random forest classifiers on each individually. The dimensionality of the Gram matrices post-reshaping is 2016, 8128, 32640, 130816, and 130816 respectively. Considered separately, the random forest classifiers built on the first three of these style representations performed better than the online linear classifier using the full style representation, and better than the baseline convolutional neural network, achieving top-1% accuracies of 27.84%, 28.97%, and 33.46% respectively. The random forest classifiers built on the latter two style representations performed considerably worse. The results of the neural style representation based models are summarized in Table II.

The dimensionality of the style representation is a significant hinderance to its effective usage, suggesting that an

TABLE II. STYLE REPRESENTATION RESULTS

Model	Accuracy (top 1%)
Full Style Representation - Linear Classifier	13.21
ReLU1_1 Random Forest	27.84
ReLU2_1 Random Forest	28.97
ReLU3_1 Random Forest	33.46
ReLU4_1 Random Forest	9.79
ReLU5_1 Random forest	10.18

inexpensive way to improve results may simply be to perform some dimensionality reduction such as principal component analysis (PCA) before passing the representation to a classification algorithm. In our experiments, in contrast to results reported in [17], we observed a significant loss in accuracy when dimensionality reduction was even lightly utilized. For instance, when PCA was used to reduce dimensionality while preserving 90% of the variance in the data from the layer ReLU1_1 style representation, the accuracy of the random forest model on that representation was reduced from 27.84% to 17%. We believe this is may be due to our use of a larger, less balanced, and less homogeneous dataset.

V. CONCLUSION AND FUTURE WORK

The neural-style representation of an artwork offers a novel approach to the algorithmic detection of artistic style that is founded on the ability to use convolutional neural networks to successfully transfer the style of one image to another. Use of the neural style representation with relatively simple classifiers such as random forests produce competitive performance with limited effort on hyperparameter search, with top 1% accuracy comparable to results presented in [9] but with a significantly larger number of artistic style categories. However, these experiments also demonstrate that a modern deep neural network, when pretrained for a vision task and finetuned for artistic style classification may obtain superior results. The best results obtained using the neural-style representation of an artwork were obtained when models suitable for high-dimensional nonlinear data were constructed individually on the first three Gram matrices that form the building blocks of the neural-style representation, while the results obtained using the full neural style representation likely could be improved with additional tuning and hyperparameter search.

Despite the aesthetically pleasing results that can be obtained using the neural-style algorithm for style transfer, it appears that the various neural-style representations described in this paper do not fully encode the art-historical definition of artistic style. Given the fuzzy definition of artistic style, we expect the irreducible error on this problem to be significant and this is to some degree expected. However, it is clear that the information contained in the neural style representation is highly relevant to the algorithmic detection of artistic style in paintings and has significant predictive ability. Better understanding the information encoded in the neural style representation will be crucial to improving on the results presented in this paper, and is a focus of future work.

Another path forward may be indicated by recent work published after the experiments detailed in this paper were conducted, in which methods are detailed designed to qualitatively improve the results obtained using the neural style algorithm for image style transfer [15], [16]. The methods detailed in these works include the use of a layer weighting scheme, the inclusion of more and different layers than those used in the experiments above, shifting activations when computing Gram matrices to eliminate sparsity, and the incorporation of correlations between layers. All of these modifications have the potential to meaningfully alter the style representation of the artwork under consideration and may be more informative. Beyond that, there have been numerous significant improvements to convolutional neural network architectures since the the development of the VGG networks used to extract style representations in this work, including the introduction of so-called inception networks, residual neural networks, wide residual neural networks, densenets, and highway networks, to name only a few [19], [25]–[28]. The incorporation of recent techniques into style representation, combined with the use of modern model architectures is also an avenue future work.

ACKNOWLEDGMENTS

The author would like to thank NVIDIA for GPU donation to support this research, Wikiart.org for providing many of the images, the website Kaggle.com for hosting the data, and Kiri Nichols for sourcing the data.

REFERENCES

[1] E. Fernie, *Art History and its Methods: A Critical Anthology*. London: Phaidon, 1995.

[2] B. Lang, *The Concept of Style*. Cornell University Press, 1987.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[4] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: http://arxiv.org/abs/1703.06870

[5] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[6] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

[7] D. Keren, "Recognizing image style and activities in video using local features and naive bayes," *Pattern Recogn. Lett.*, vol. 24, no. 16, pp. 2913–2922, Dec 2003. [Online]. Available: http://dx.doi.org/10.1016/S0167-8655(03)00152-1

[8] L. Shamir, T. Macura, N. Orlov, D. M. Eckley, and I. G. Goldberg, "Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art," *ACM Trans. Appl. Percept.*, vol. 7, no. 2, pp. 8:1–8:17, feb 2010. [Online]. Available: http://doi.acm.org/10.1145/1670671.1670672

[9] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," in *International Conference on Data Mining Workshops*. IEEE, 2015.

[10] S. Karayev, A. Hertzmann, H. Winnemoeller, A. Agarwala, and T. Darrell, "Recognizing image style," *CoRR*, vol. abs/1311.3715, 2013. [Online]. Available: http://arxiv.org/abs/1311.3715

[11] Y. Li, J. Yosinski, J. Clune, H. Lipson, and J. E. Hopcroft, "Convergent learning: Do different neural networks learn the same representations?" 2015, pp. 196–212.

[12] L. Wang, L. Hu, J. Gu, Z. Hu, Y. Wu, K. He, and J. Hopcroft, "Towards understanding learning representations: To what extent do different neural networks learn the same representation," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 9607–9616.

[13] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[14] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, vol. 2, no. 11, p. e7, 2017.

[15] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos," *arXiv preprint arXiv:1604.08610*, 2016.

[16] R. Novak and Y. Nikulin, "Improving the neural algorithm of artistic style," *CoRR*, vol. abs/1605.04603, 2016. [Online]. Available: http://arxiv.org/abs/1605.04603

[17] S. Matsuo and K. Yanai, "Cnn-based style vector for style image retrieval," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '16. New York, NY, USA: ACM, 2016, pp. 309–312. [Online]. Available: http://doi.acm.org/10.1145/2911996.2912057

[18] B. Graham, "Fractional max-pooling," *CoRR*, vol. abs/1412.6071, 2014. [Online]. Available: http://arxiv.org/abs/1412.6071

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[20] ——, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[21] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2377–2385.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015. [Online]. Available: http://arxiv.org/abs/1409.4842

[26] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2261–2269.

[28] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.